Laboration 2

SF1935 Sannolikhetsteori och statistik med tillämpning inom maskininlärning

Pontus Linde, ponlinde@kth.se
Leon Fällman, leonfa@kth.se

2023-05-15

Förberedelseuppgifter

1.

a) Bestäm ML-Skattningen av b.

$$f_X(x) = \frac{x}{b^2} e^{\frac{-x^2}{2b^2}}$$

Skapa Likelihoodfunktionen

$$L(b) = f_{X_1,X_2,\ldots,X_n}(x_1, x_2, x_3) = \{\text{oberoende}\} = f_{X_1}(x_1) f_{X_2}(x_2) \ldots f_{X_n}(x_n) =$$

$$= \frac{x_1}{b^2} e^{\frac{-x_1^2}{2b^2}} \frac{x_2}{b^2} e^{\frac{-x_2^2}{2b^2}} \ldots \frac{x_n}{b^2} e^{\frac{-x_n^2}{2b^2}} = \frac{x_1 x_2 \ldots x_n}{b^{2n}} e^{-\frac{1}{2b^2}\sum_{i=1}^n x_i^2}$$

Skapa logLikelihoodfunktionen:

$$\ln(L(b)) = \ln\left(\frac{x_1 x_2 \ldots x_n}{b^{2n}} e^{-\frac{1}{2b^2}\sum_{i=1}^n x_i^2}\right) = -2\ln(b) + \left(\sum_{i=1}^n \ln(x_i)\right) - \frac{1}{2b^2}\left(\sum_{i=1}^n x_i^2\right)$$

ML-skattningen för b bestäms som det b som löser:

$$0 = \frac{d}{db}\ln(L(b)) = \frac{d}{db}\left[-2\ln(b) + \left(\sum_{i=1}^n \ln(x_i)\right) - \frac{1}{2b^2}\left(\sum_{i=1}^n x_i^2\right)\right] = -\frac{2n}{b} + \frac{1}{b^3}\sum_{i=1}^n x_i^2$$

Vi får då:

$$\frac{2n}{b} = \frac{1}{b^3}\sum_{i=1}^n x_i^2$$

$$2nb^2 = \sum_{i=1}^n x_i^2$$

$$b^2 = \frac{1}{2n}\sum_{i=1}^n x_i^2$$

$$b = \sqrt{\frac{1}{2n}\sum_{i=1}^n x_i^2}$$

Dvs, ML-skattningen för b är:

$$b^*_{\text{obs}} = \sqrt{\frac{1}{2n}\sum_{i=1}^n x_i^2}$$

b) Bestäm MK-skattningen av b.

Vi börjar med att bestämma väntvärdet:

$$E(X_i) = \int_{-\infty}^{\infty} x \frac{x}{b^2} e^{\frac{-x^2}{2b^2}} = \frac{1}{b^2} \int_{-\infty}^{\infty} x^2 e^{\frac{-x^2}{2b^2}} = \ldots = b\sqrt{\frac{\pi}{2}}$$

Vi bildar funktionen Q(b). MK-skattningen av b kommer vara det värde på b som minimerar Q(b).

$$Q(b) = \sum_{i=1}^{n} \left[x_i - E_i(b)\right]^2 = \{\text{alla } E_i(b) \text{ är samma}\} = \sum_{i=1}^{n} \left[x_i - E(b)\right]^2 =$$

$$= \sum_{i=1}^{n} \left[x_i - b\sqrt{\frac{\pi}{2}}\right]^2$$

$$Q^{'}(b) = 0$$

$$n\pi b - \sqrt{2\pi} \sum_{i=1}^{n} x_i = 0$$

$$\sqrt{2\pi} \sum_{i=1}^{n} x_i = n\pi b$$

$$\frac{1}{n} \sum_{i=1}^{n} x_i = \frac{\pi}{\sqrt{2\pi}} b$$

$$b = \frac{\frac{1}{n} \sum_{i=1}^{n} x_i}{\sqrt{\frac{\pi}{2}}}$$

MK-Skattningen av b blir således:

$$b^*_{\text{obs}} = \frac{\frac{1}{n} \sum_{i=1}^{n} x_i}{\sqrt{\frac{\pi}{2}}} = \frac{\overline{x}}{\sqrt{\frac{\pi}{2}}}$$

2.

Vi kan säga att den är approximativt normalfördelad om vi använder MK-skattningen av b eftersom MK-skattningen endast är en summa av många oberoende likafördelade stokastiska variabler. Därav kan man enligt FS.12.3 säga att:

$$I_b = b^*_{\text{obs}} \pm D^*_{\text{obs}} \lambda_{\frac{\alpha}{2}}$$

$$b^*_{\text{obs}} = \frac{\overline{x}}{\sqrt{\frac{\pi}{2}}}$$

$$b^* = \frac{\overline{X}}{\sqrt{\frac{\pi}{2}}}$$

$$\lambda_{\frac{\alpha}{2}} = 1.96 \text{ (för en } 95\% \text{ konfidensgrad)}$$

$$V(b^*) = V\left(\frac{\overline{X}}{\sqrt{\frac{\pi}{2}}}\right) = \frac{2}{\pi}V(\overline{X}) = \{\text{oberoende}\} = \frac{2}{\pi n^2}(V(X_1) + V(X_2) + ... + V(X_n)) =$$

$$= \left\{V(X_i) = \frac{4-\pi}{2}b^2\right\} = \frac{2}{\pi n^2}\sum_{i=1}^{n}\frac{4-\pi}{2}b^2 = \frac{1}{\pi n^2}\sum_{i=1}^{n}(4-\pi)b^2 = \frac{1}{\pi n}(4-\pi)b^2$$

$$D(b^*) = \sqrt{V(b^*)} = \sqrt{\frac{1}{\pi n}(4-\pi)b^2}$$

$$D^*_{\text{obs}} = D^*(b^*)_{\text{obs}} = \sqrt{\frac{1}{\pi n}(4-\pi)(b^*_{\text{obs}})^2} = \sqrt{\frac{1}{\pi n}(4-\pi)\left(\frac{\overline{x}}{\sqrt{\frac{\pi}{2}}}\right)^2} = \frac{\overline{x}}{\sqrt{\frac{\pi}{2}}}\sqrt{\frac{1}{\pi n}(4-\pi)}$$

Vi får då:

$$I_b = b^*_{\text{obs}} \pm D^*_{\text{obs}}\lambda_{\frac{\alpha}{2}} = \frac{\overline{x}}{\sqrt{\frac{\pi}{2}}} \pm \frac{\overline{x}}{\sqrt{\frac{\pi}{2}}}\sqrt{\frac{1}{\pi n}(4-\pi)}\ 1.96$$

3.

Idén bakom linjär regression är att kunna skatta värdet på en variabel med hjälp av värden på en andra variabler. För att göra en linjär regression i MATLAB med hjälp av funktionen `regress` vill vi skapa en matris:

$$\text{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

För att sedan lösa det överbestämda systemet

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Där vårt resultat blir den linjära regressionen för β_0 och β_1 Syntaxen i MATLAB blir

```
beta_hat = regress(log(y), X)
```

Där log(y) är vårt faktiska värde för ekvationen.

Laborationsuppgifter

Problem 1 - Simulering av konfidensintervall

a) Hur många av dessa intervall kan förväntas innehålla det sanna värdet på μ?

Ett konfidensintervall med konfidensgraden 1 - α innebär att för vårt intervall beräknas mätningarna övertäcka μ (mu i kodexemplet) med den givna sannolikheten 1 - α. I detta scenario innebär det att konfidensgraden är 95 (1 - 0.05) vilket innebär ett 95% konfidensintervall. Detta medför att för 100 mätningar, beräknas 95 utav dessa övertäcka parametern μ. Att ett intervall övertäcker det sanna värdet på μ menas alltså att intervallet innehåller även samma värde.

b) Vad visar de horisontella strecken och det vertikala strecket?

Det vertikala strecket (grönt i MatLab-körningen) symboliserar vårt sanna värde på parameter μ. Den är konstant i varje körning för att vi tilldelar parametern värdet 2.

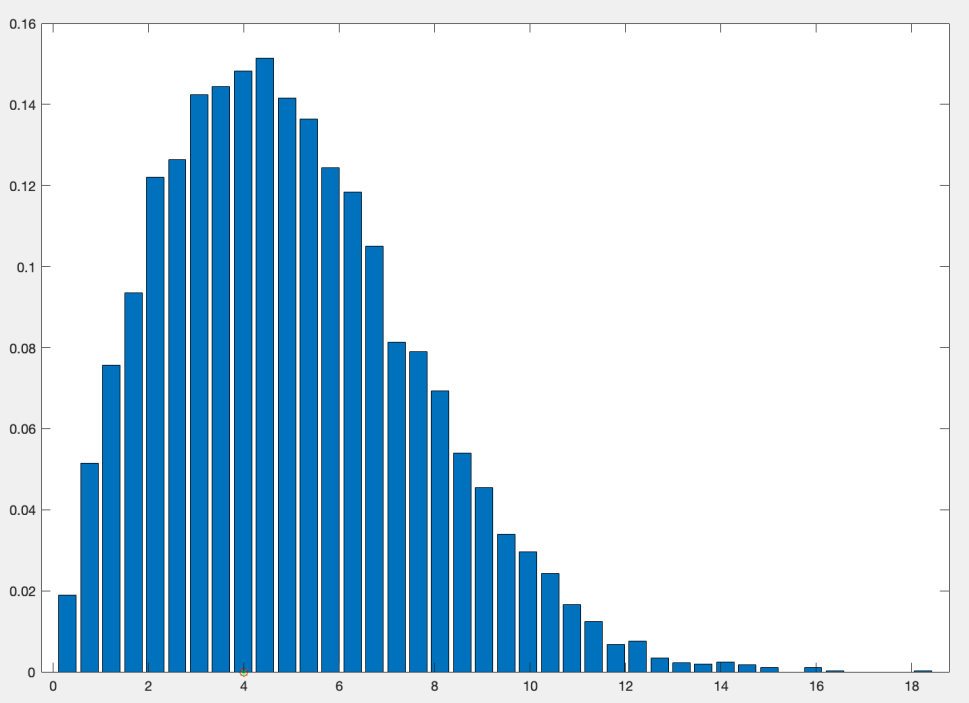
De horisontella strecken är varje konfidensintervall (100 stycken) där varje blått streck är ett intervall som innehöll μ och där varje rött streck är ett intervall som inte innehöll μ.

c) Hur många av de 100 intervallen innehåller det sanna värdet på μ? Stämmer resultatet med dina förväntningar? Kör simuleringarna flera gånger.

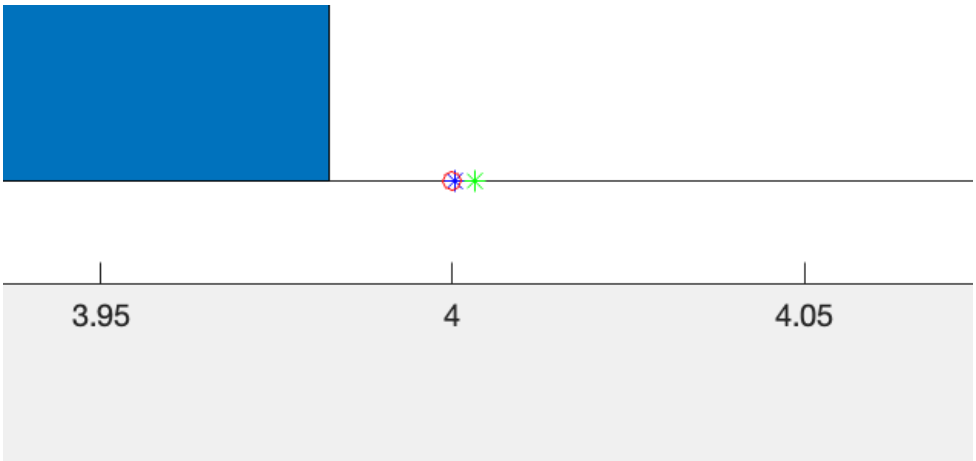
Det verkar stämma överens med förväntningarna. Genom att köra koden 5 gånger erhölls resultaten nedan:

1. 92 träff, 8 miss
2. 93 träff, 7 miss
3. 95 träff, 5 miss
4. 99 träff, 1 miss
5. 97 träff, 3 miss

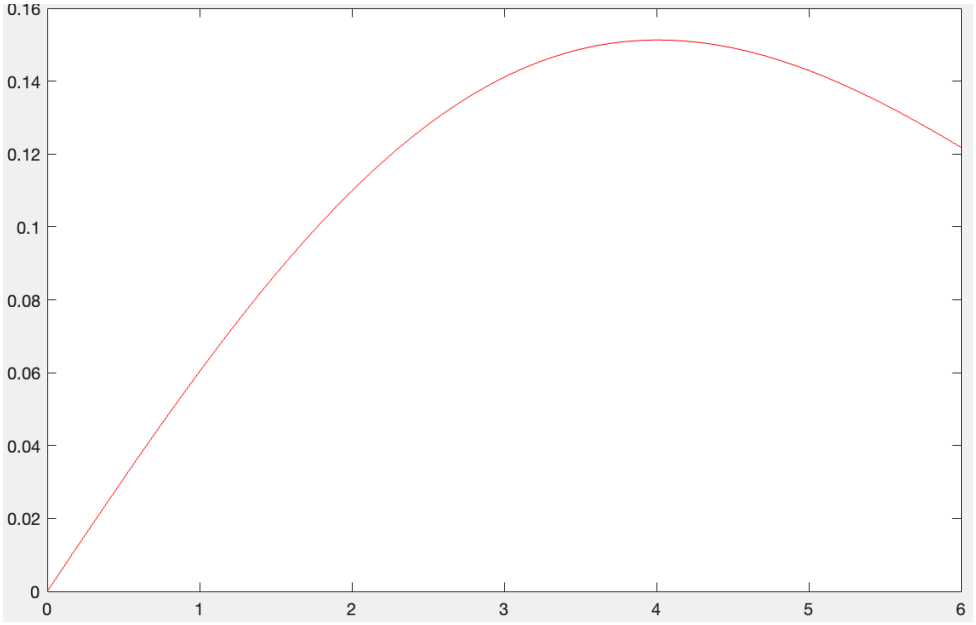
Problem 2 - Maximum likelihoodskattning och minsta kvadrat- skattning



Figur 1 - Plot av Rayleighfördelningen



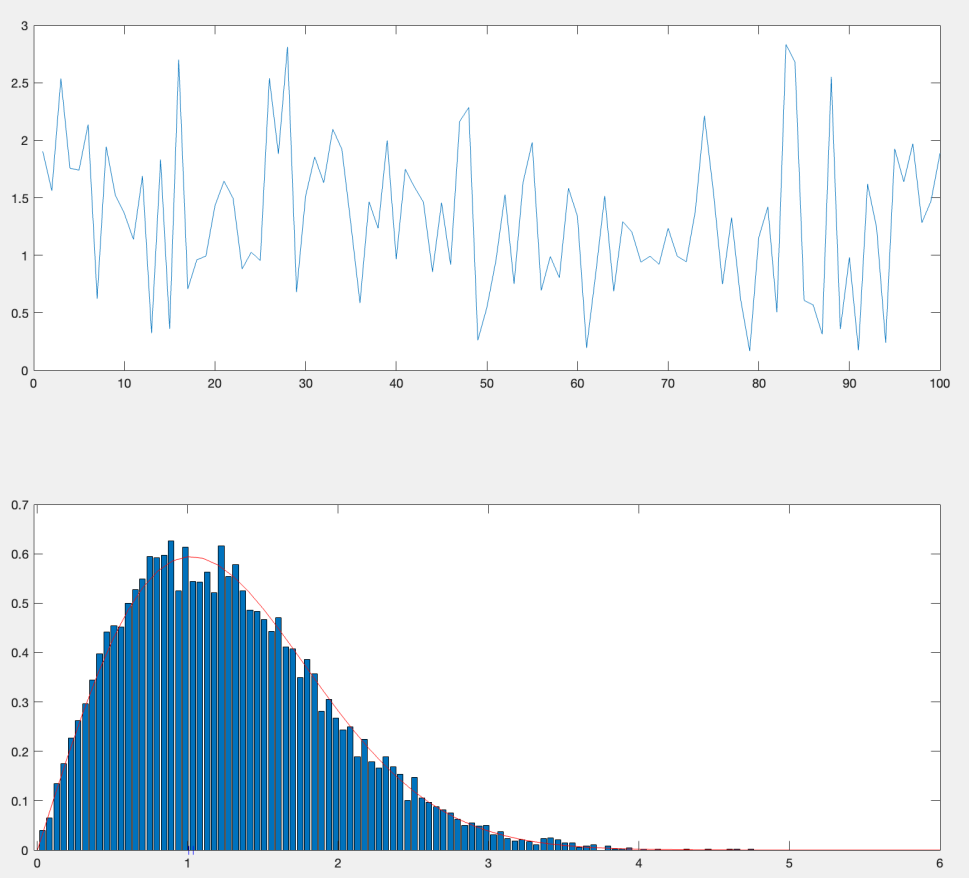
Figur 2 - Inzoomad plot av Rayleighfördelningen



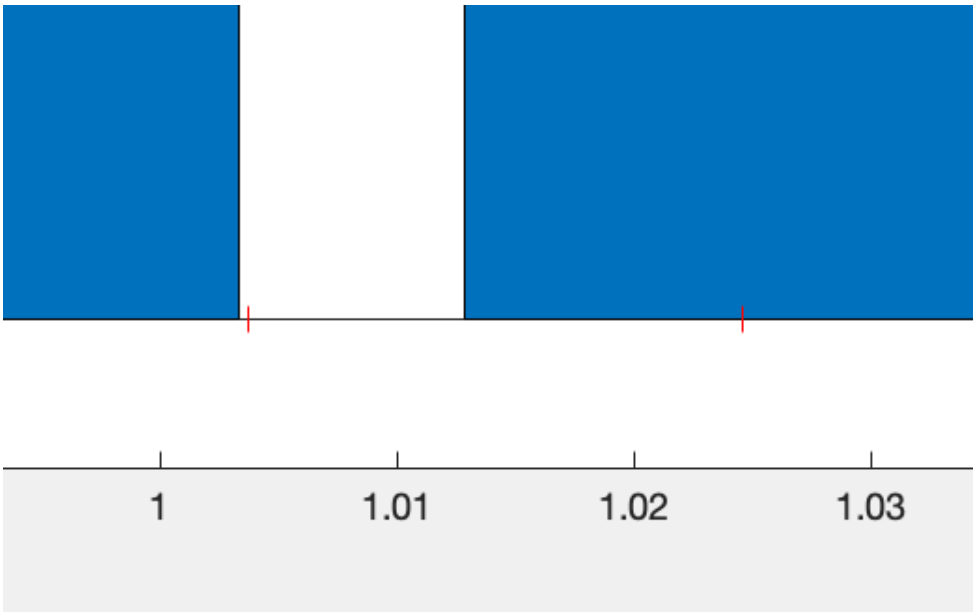
Figur 3 - Plot av täthetsfunktionen med ML-skattning

Den röda ringen i *Figur 1* representerar värdet på b, den blåa asteriksen är ML-skattningen av b och den gröna asteriksen är MK-skattningen av b. I *Figur 2* syns dessa värden tydligare. *Figur 3* visar även ploten av täthetsfunktionen med ML-skattningen.

Problem 3 - Konfidensintervall för Rayleighfördelning



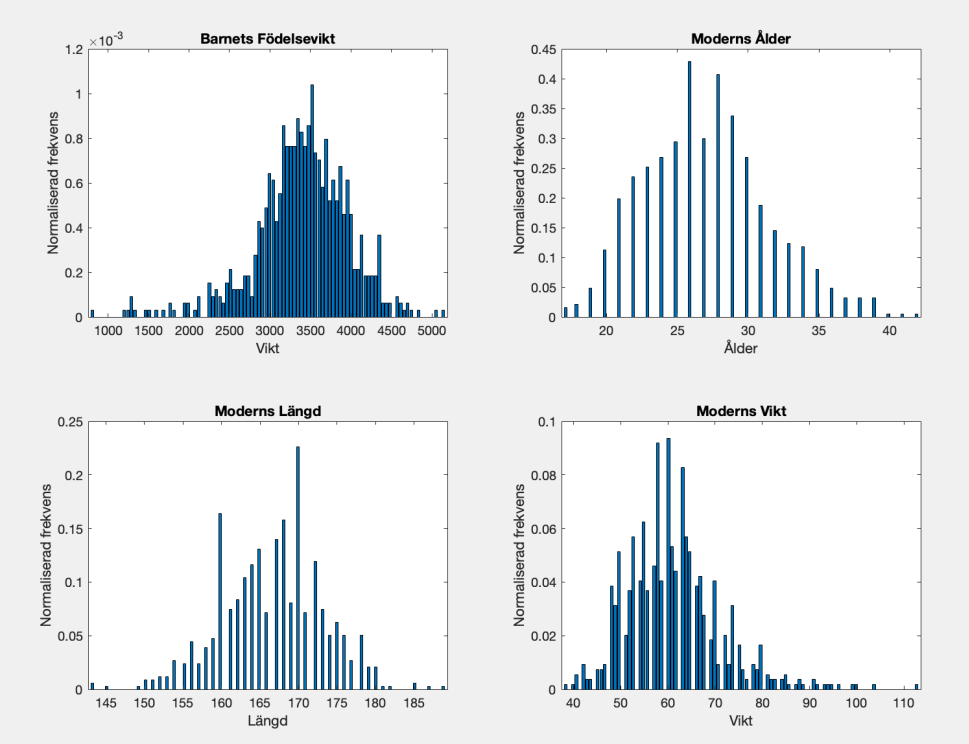
Figur 4 - Signalen och täthetsfunktionen



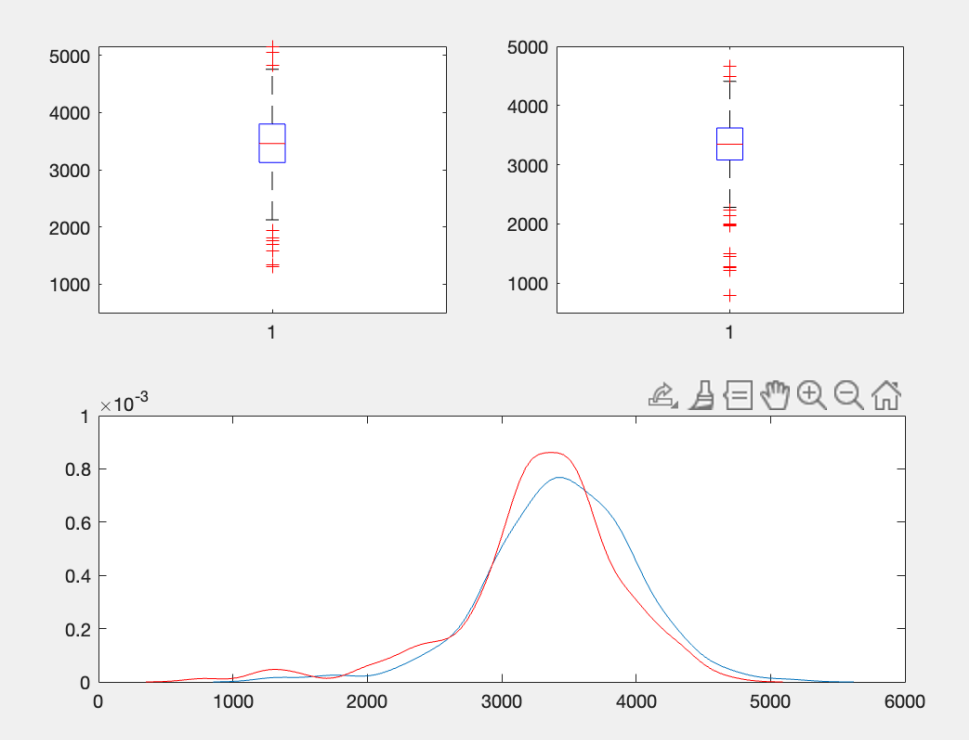
Figur 5 - 95% approximativt konfidensintervall för parametern b

Fördelningen ser ut att passa bra.

Problem 4 - Jämförelse av fördelningar hos olika populationer



Figur 6 - Fyra olika histogram som visar fördelningarna för barnets födelsevikt, moderns ålder, moderns längd respektive moderns vikt.

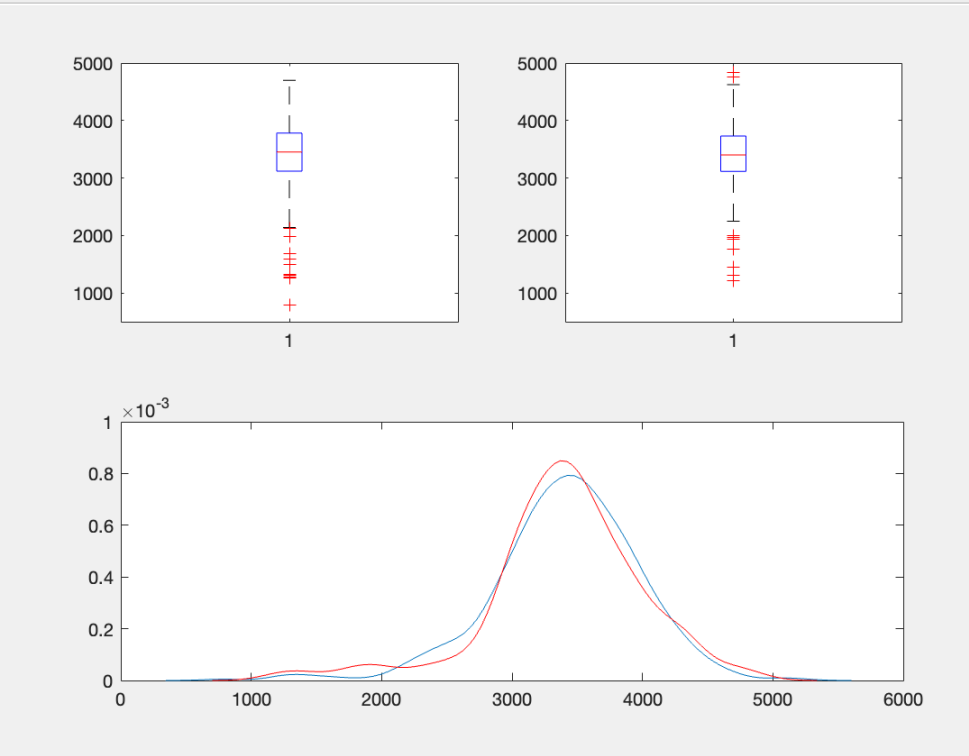


Figur 7 - Första boxogrammet till vänster är för mödrar som inte röker eller har slutat, boxogrammet till höger är för mödrar som röker, tredje linjediagrammet är en jämförelse mellan de två

En intressant anmärkning är att för mödrar som inte röker eller har slutat, är minimum värdet för barnets vikt 1310 g, gentemot mödrar som röker registrerades ett minimum för barnets vikt på 785. På samma sätt för mödrar som slutat eller inte röker registrerades ett maximum för barnets vikt 5160 g gentemot 4660 g för mödrar som röker.

Kan även se att för förstnämnda grupp av mödrar så har erhålls en median på 3460 g för barnets vikt gentemot en median på 3345 g för andra testgruppen.

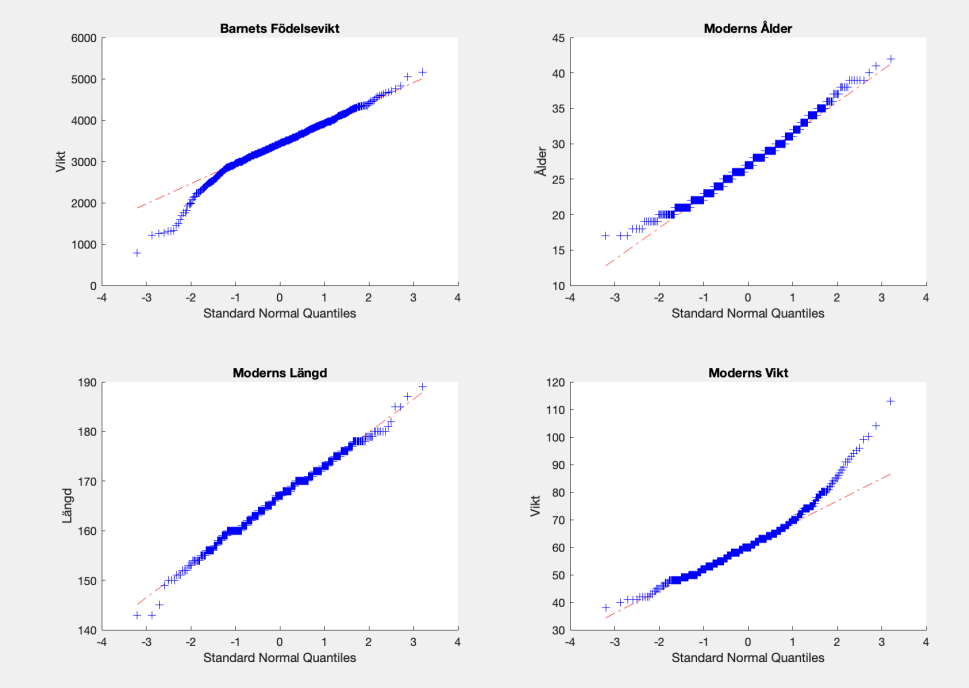
En slutsats vi kan dra är barn vars moder har rökt, löper större risk för att bli född underviktig till extremt underviktig, genom att se i fördelningsfunktionerna täcker den röda kurvan (mödrar som röker) mer datapunkter för de lägre vikterna (den vänstra kvantilen).



Figur 8 - Första boxogrammet till vänster är för mödrar som inte dricker eller har slutat pga graviditet, boxogrammet till höger är för mödrar som dricker, tredje linjediagrammet är en jämförelse mellan de två

Vi kan se att på samma sätt som för sambandet mellan barnets födelsevikt och rökande kvinnor att det finns ett samband mellan drickande mödrar och låg födelsevikt för barnet. En slutsats som går att dra från fördelningarna är att för barn med mödrar som dricker under graviditeten löper en större risk för att födas med lägre vikt. Vi ser även i *Figur 8* att det är fler barn med mödrar som dricker som har lägre födelsevikt, samt att kurvan är förskjuten mer mot undervikt.

Problem 5 - Test av normalitet

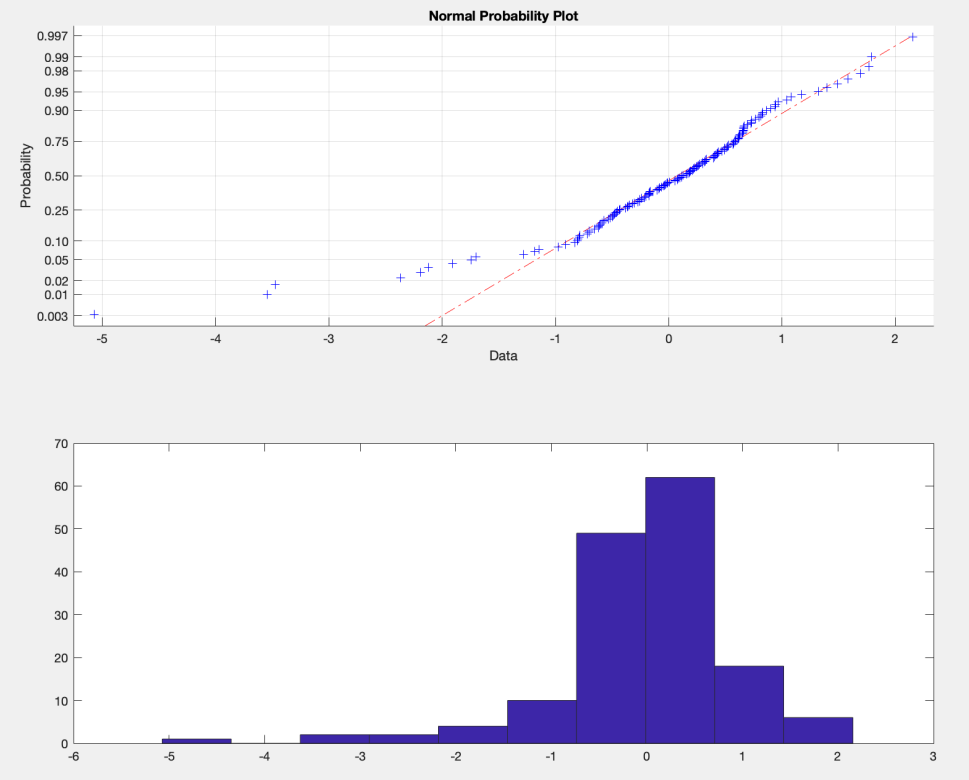


Figur 9 - jämförelse av den empiriska datamängdens kvantiler med kvantilerna för en normalfördelning mha. qqplot

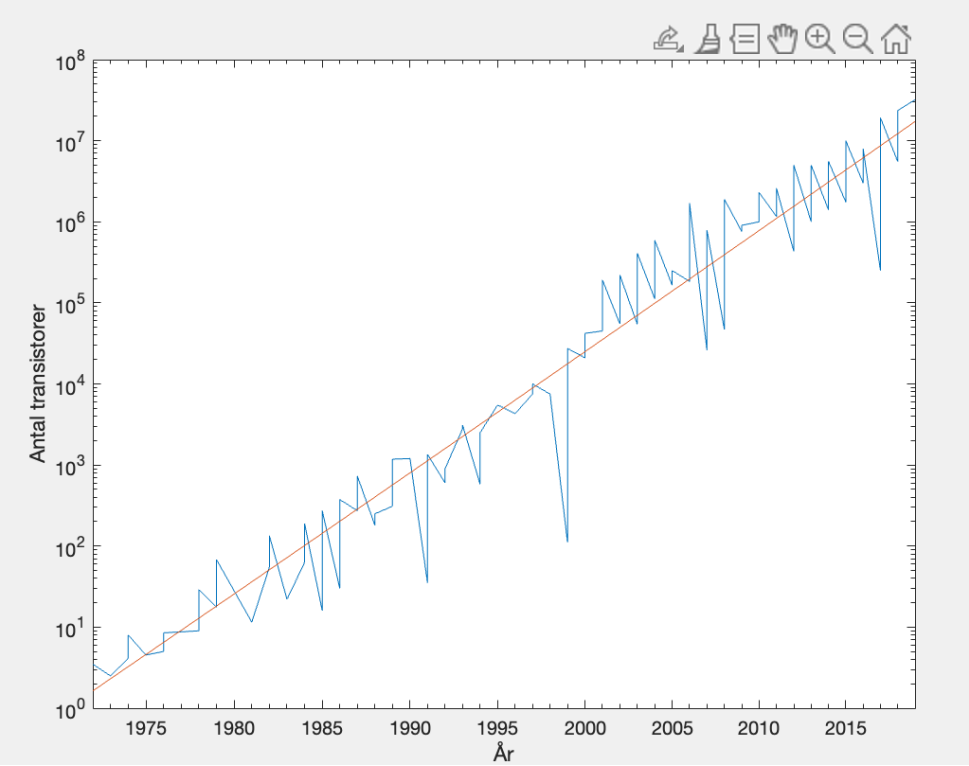
Att jbtest returnerar 1 innebär att nollhypotesen avvisas med givna signifkantsnivå, 0 innebär att den ej avvisas för signifikantsnivån.

Med användning av `jbtest` ser vi att den returnerar 1 för alla parametrar förutom för moderns längd returnerar funktionen 0 för signifikantsnivån 5%. Att funktionen returnerar 1 innebär att med 95% säkerhet kan vi anta att fördelningarna inte är normalfördelade. För moderns längd, då funktionen returnerar 0, kan vi anta med 95% säkerhet att fördelningen inte är icke-normalfördelad. Således kan vi enligt Jarque-Berus test anta att moderns längd är normalfördelad.

Problem 6 - Enkel linjär regression



Figur 10 -



Figur 11 - En loglog plot av den faktiska datan mot en den linjära regressionen för datat

Genom att kika på det nedre diagrammet, ser fördelningen att komma ifrån en normalfördelning (nästan symmetrisk kring 0). Anledningen som vi tror varför inte den är perfekt normalfördelad är för att vi saknar tillräckligt många datapunkter.

R2 värdet hämtas från regress och vi får då

$$\mathrm{R2} = 0.9586$$

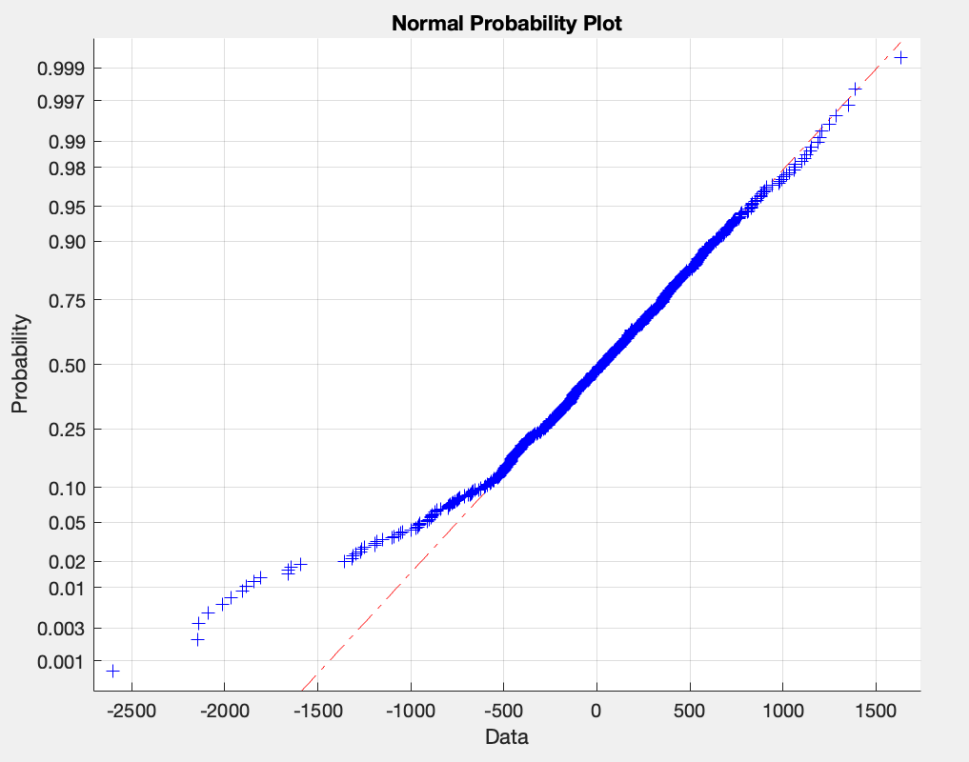
Genom att använda oss av modellen kan vi göra en approximation för värdet 2025. Detta görs på följande sätt och får då svaret:

```
y2025 = exp(1)^(beta_hat(1) + beta_hat(2)*2025);
```

$$\mathrm{y2025} = 1.3599e+08$$

Vilket är ett rimligt svar om man jämför med de tidigare värderna.

Problem 7 - Multipel linjär regression



Figur 12 - Plot av residualerna

För residualerna kan vi se att man i vissa fall underskattas barnets födelsevikt (svansen till vänster) oftare än vad man överskattar den.

Vi kan även se att konfidensintervallen givna från regress blir:

1.0e+03 ∗

Konstanter: 2.5137 3.0290
ModerVikt: 0.0072 0.0154
ModerRökare: -0.2402 -0.0640
ModerAlkoholist: -0.1304 0.0515

Vi ser att för både ModerVikt och ModerRökare så innefattar inte det 95% konfidensintervallet 0, vilket visar på att dessa två parametrar har en signifikant påverkan på barnets födelsevikt. Däremot för ModerAlkoholist så innefattar intervallet 0, vilket gör att vi inte med säkerhet kan säga om denna parameter påverkar barnets födelsevikt.